INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Investigating the Visual Lombard Effect with Gabor Based Features

Waito Chiu[1], *Yan Xu*[1], *Andrew Abel*[1] ,*Chun Lin*[2], *Zhengzheng Tu*[2]

[1] Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[2] School of Computer Science and Technology, Anhui University, Hefei, 230601, China

Waito.Chiu18@student.xjtlu.edu.cn; yan.xu@xjtlu.edu.cn; andrew.abel@xjtlu.edu.cn; lc20191001@163.com; zhengzhengahu@163.com

Abstract

The Lombard Effect shows that speakers increase their vocal effort in the presence of noise, and research into acoustic speech, has demonstrated varying effects, depending on the noise level and speaker, with several differences, including timing and vocal effort. Research also identified several differences, including between gender, and noise type. However, most research has focused on the audio domain, with very limited focus on the visual effect. This paper presents a detailed study of the visual Lombard Effect, using a pilot Lombard Speech corpus developed for our needs, and a recently developed Gabor based lip feature extraction approach. Using Kernel Density Estimation, we identify clear differences between genders, and also show that speakers handle different noise types differently.

Index Terms: Lombard Effect, Gabor Features, Lip Features

1. Introduction

Noise is common in many speech environments and can affect communication between speakers and listeners. To cope, speakers tend to change their speech patterns to be understood. The increased vocal effort made in noisy environments is known as the Lombard Effect [1] [2], and these features can be classified as acoustic and visual. As well as vocal effort, other characteristics include vowel duration and fundamental frequency (F0) [3]. Lane and Tranel [4] reported that in noisy environments, there is an increase in pitch, speaking rate and speech intensity. Garnier et al. [5] found that speakers have different communication strategies in different noise environments. Bradlow et al. [6] studied 20 speakers and reported that female speakers were more intelligible than male. Another study compared speech produced in noise and silence and found that speech in noise was more intelligible [7]. Other research has identified changes at the Phonetic level [8] [7]. Certain types of noise are also more disruptive than others [9]. Garnier et al. [5] reported increases in word duration and voice intensity in white noise, and found that white noise generates larger increases than cocktail party noise. We can therefore conclude that there is a relationship in speech characteristic changes with regard to noise and gender.

However, although articulatory variation in acoustic Lombard speech is well studied, there is relatively limited research with regard to measuring the change in visual properties like face, head, mouth and lip motion. Of the limited studies identified, Fitzpatrick et al. found that when participants communicated face-to-face in noise, talkers increased their lip-area and lip-width more compared to not seeing listener's faces [3]. Kim et al. found that the motion of speakers' mouth and jaw was larger when speaking in noise [4]. Garnier et al. found that speakers produced shorter protruded lip pinching and wider swallowed lip pinching in white noise [5]. Heracleous et al. used Kernel Density Estimation to simulate a visual speech automatic detector. The model recognised higher values in the outer lip-height in noisy speech [10]. However, these are limited studies, with much further quantification required.

This visual change is a pertinent topic because many audio-visual noise removal or recognition systems are evaluated by adding noise to the audio signal [11] [12], but the visual signal is usually left unmodified. However, as discussed above, speakers change their speech in the presence of noise, and so realistically, we would expect to see some modifications to the visual signal. If a recognition system has been trained on Lombard speech, then statistical models of Lombard speech may improve recognition results.

In this paper we use a Gabor based feature extraction [13] method recently developed by the authors to extract precise geometric features, and introduce a new audiovisual Lombard speech Database, which contains volunteers of different genders speaking in different speech environments (with different noise types and levels). We use Kernal Density Estimation (KDE) to visualize changes in visual speech by comparing different scenarios, categorized by gender, type of noise, and noise level. By studying these features in depth, we aim to identify differences in different scenarios, and the level of effect.

2. Audiovisual Lombard Speech Database

It is difficult to record accurate and precise visual Lombard speech, because participants may manually overcompensate, or deliberately disregard the effect. This means that creating a Lombard Corpus that can be used for comparison is very challenging. Here, we present a small database, suitable for the pilot study reported in this work.

2.1. Subjects

Fifteen participants, nine females and six males were recorded. Thirteen of them were native English speakers and were university graduates, recruited at the University of Stirling, Scotland. In this work, we focused on data from three females and three males, because they were the best quality recordings, with a good distance from the camera, and clear articulation in all cases, suitable for good quality feature extraction. The datasets were categorized by their genders and noise level for this analysis. Each speaker was asked to read 10 TIMIT sentences.

2.2. Procedure

All recording took place in a booth with individual speakers directly facing a camera. However, to make the task seem more natural, the booth was designed with mirrors so that the camera was not visible, and the volunteer was looking directly at

http://dx.doi.org/10.21437/Interspeech.2020-1291

the experiment administrator. Participants were invited to read a list of ten sentences without headphones on. They then read the same sentences with white noise played (see Fig. 1) while wearing noise cancelling headphones. This was repeated at four different levels (60dB, 66dB, 72dB and 78dB), and with conversational babble noise. Participants were given 1 minute breaks between different noise levels.

Figure 1: *Example speakers from the Lombard Speech database, showing male (bottom), and female(top).*

2.3. Noise Levels and Types

The audio signal was digitized at a rate of 48kHz and over 32bits. Two noisy environments were used: conversational babble (to simulate an multiple speaker environment) and white noise, played through noise cancelling headphones and kept below 78dB. Speakers were first recorded in a silent condition as a reference, and then in both noisy environments. Noise was not recorded, as the noise cancelling headphones prevented sound escaping.

2.4. Dataset Extraction

The 10 chosen Timit sentences are listed below. Fifteen keywords (underlined) with the targeted vowels (/i:/, /ae/, /u:/) were chosen. These vowels give the maximum/ minimum of the frequency, in term of their position and shape, to plot a vowel space chart. These vowels have varying positions and shapes. Vowel /i:/ is closed (high) but vowel /ae/ is open (low), while /i:/ has tongue at the front but /u:/ has the tongue position at the back. Individual keywords were extracted from the video files, and precise visual features extracted.

1. She had your dark suit in greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Do they make class-biased decisions?
4. He took his mask from his forehead and threw it, unexpectedly, across the desk.
5. Make lid for sugar bowl, the same as jar lids, omitting design disk.
6. The clumsy customer spilled some expensive perfume.
7. The viewpoint overlooked the ocean.
8. Please dig my potatoes up before frost.
9. I'd ride the subway but I haven't enough change.
10. Grandmother outgrew her upbringing in petticoats.

3. Visual Feature Extraction

To extract precise 3D geometric features, we use a newly developed lightweight lip feature extraction system, as developed by the authors and introduced in previous research [13] [14]. This uses Gabor based transforms to extract the mouth region. This allows us to see precise changes in geometric features, including the lip height, lip width, and lip area, based on psychological research into human faces [15] . The flow chart of the process is shown in Fig. 2. Due to space limitations, more detail about system performance can be found in Xu et al. [13], but there are several key lip feature extraction steps:

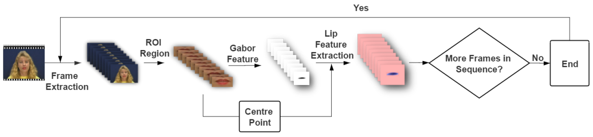

Figure 2: *Key lip feature extraction system stages*

1. Extract individual frames from video and extract ROI using the Dlib toolkit [16], returning the mouth region, as well as the central point of the mouth $P(x, y)$. A trained model file, the Shape-Predictor-68-Face-Landmarks (see Fig. 3.) was used [17]. For each frame, 68 landmark features were detected by the dlib toolkit and 4 features were chosen to identify the mouth region (ROI) and centre point, as shown in Fig. 3.

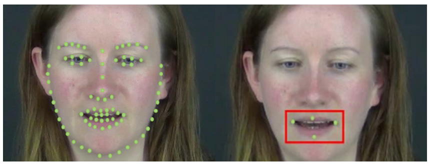

Figure 3: *(left)Landmark features detected by the dlib toolkit, and (right) ROI extracted from 4 landmark features.*

2. Obtain Gabor features. A 2D Gabor filter is used for edge detection and extraction of texture features of greyscale images. The Gabor transform is performed as follows:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{x\prime^2 + \gamma^2 y\prime^2}{2\sigma^2}) \cos(2\pi\frac{x\prime}{\lambda} + \psi) \quad (1)$$

where:

$$x\prime = x\cos\theta + y\sin\theta$$
$$y\prime = -x\sin\theta + y\cos\theta$$

This is implemented using the Python function:

$$cv2.getGaborKernel((Ksize, Ksize), \sigma, \theta, \lambda, \gamma, \psi) \quad (2)$$

This has six parameters:

$Ksize$: The size of the Gabor kernel.
σ: The standard deviation of the Gaussian function used in the Gabor filter.
θ: The orientation of the normal to the parallel stripes of the Gabor function. In our paper, all the orientation in paper are 90 degrees.
λ: The wavelength of the sinusoidal factor in the equation.
γ: The spatial aspect ratio.
ψ: The phase offset. it is defined as 0 by default.

In this work, four parameters need to be adjusted: $ksize$, σ, λ, and γ. Preliminary investigation identified the optimal parameters as: $ksize$ = 12, σ = 5, λ = 15 and γ = 0.5.

3. Identify the target pixels. The Python function $filters.threshold - yen$ is used to determine the optimal segmentation threshold using the Yen algorithm [18]. This identifies each pixel in the transformed image as being part of the target or background region, as shown in Fig.4. The "high" threshold based method returned the most accentuated lip features.

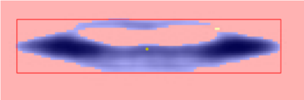

Figure 4: *Gabor Transformed ROI.*

4. Select the mouth region according to the $P(x; y)$ and return 7 parameters, as discussed in [13]: width, height, area, intensity, x and y values of central point, and orientation. The box width is the inter-lip width. The height is the inter-lip height. The area is the number of pixel, the intensity is the sum of each pixel density value (the darker the inter-lip area, the deeper the mouth opening and the larger the sound intensity), as shown in Fig.4, and Fig.5.

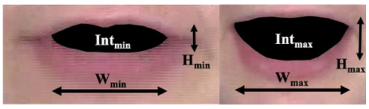

Figure 5: *Example of mouth ROI, showing height, width, area, and intensity*

5. The variance of the amplitude (maximum and minimum) of articulatory movements was calculated (e.g. Variance of width = Wmax – Wmin, see Fig.5). The variance was chosen because it removes synchronization issues in comparing parameters. For example, some vowel sound peaks are concentrated at the front (e.g. ocean) and some at the middle (e.g. please). This also minimises the effect of the background features, such as the distance and angles of the camera, illumination etc. The parameters were recorded in a column vector in ascending order, then the maximum and minimum functions return the respective values. This method specifically calculates the width, height, area and Intensity for each single word, classified under three vowels (/ae/, /i:/, /u:/) and five noise scenarios (no noise, babble noise 60-66 dB, 72-78 dB, and white noise 60-66 dB, 72-78 dB).

4. Evaluation with Kernel Density Estimation

Adriana et al. [19] reviewed 138 research papers on the topic of Automatic lip-reading (ALR) systems and found that 99 researchers applied Hidden Markov Models (HMMs) as the classifier. This approach considers that each of the emitting states in the model will be described by a continuous density distribution. Piccardi et al. [20] applied a kernel density estimation (KDE) to model emission probabilities as an input of HMMs and found that this improved HMMs accuracy in detecting human features in videos such as speed and positions. To evaluate how lip features may affect visual speech detection in this paper, we applied the KDE approach to simulate the HMM normalization effect at an early stage.

4.1. Density Function Implementation

Each function has an input of 75 tokens, each of which represents one word segment at one noise level. Therefore, each noise level has 25 data points and KDE is used to form a standard normal distribution, with tokens divided into bins. As a result, the highest probability forms a peak value, which can be compared.

Fig. 6 shows an example of the normalized inter-lip height in silence and two levels of babble for male speakers, showing mouth region height. The largest values in this KDE model are observed where the peak height is 10 pixels in silence compared to 17 pixels in babble at 60-68dB, and 19 pixels in babble at 72-78dB. These are counted as two positive shifts in results in the next section, i.e. a positive change in the babble noise condition from the no noise condition. This approach allows for observing trends in differences between speakers and environments.

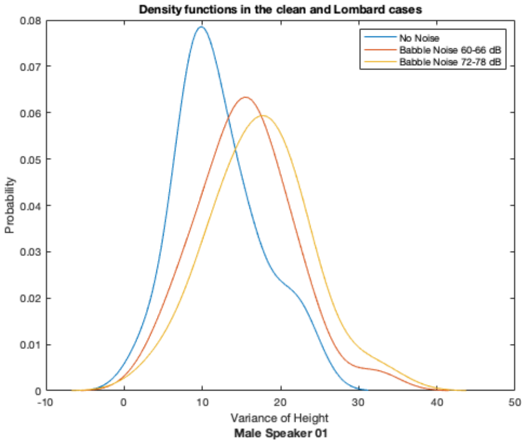

Figure 6: *A sample result of inter-lip height of a male speaker at different levels of babble noise.*

5. Results and Discussion

Here, we report on three key findings using the KDE Approach outlined above: differences between genders, noise types and noise levels, .

5.1. Evaluation of Genders

Table 1 presents parameter changes grouped by genders for all noise types. There are five parameters and each includes 75 tokens for each gender, with 5 tokens forming a group. 3 groups (i.e. 3 speakers) were tokens with no noise and 9 groups were tokens with Babble and White Noise from 60-78dB. The probabilities were counted by comparing the groups visually using KDE (see Fig. 6). A peak value increase was counted as positive, and a decrease was negative. In total, for each comparison, 12 probabilities were considered. Patterns are highlighted in the table. Male speakers were generally found to have a positive change in noisy environments (i.e. an increase). In terms of probability, mouth height has a probability of 66.67% of increasing in noise, area has a probability of 91.67%, and intensity has a probability of 100%. These findings are in line with results by Fitzpatrick et al. [3].

Conversely, female speakers were found to generally show negative changes in noisy environments. Lip width (100%), area (66.67%) and intensity (75%) have the most probability of a decrease in noise. Duration also has a probability of 66.67

percent of being shorter in noise. This result contradicts results by Maeva et al. [21]. There are a number of possible reasons for this, such as the male speakers responding differently to experimental conditions, or this being a small pilot study with a limited sample size, but nonetheless, clear gender differences were found.

Table 1: *Parameter changes for male and female speakers in babble, white noise, and silence, at all noise levels, showing grouped tokens and change in KDE, with majority changes highlighted.*

Parameter	Probability Change			
	Male		Female	
	Positive	Negative	Positive	Negative
Duration	50%	50%	33.33%	66.67%
Width	50%	50%	0%	100%
Height	66.67%	33.33%	50%	50%
Area	91.67%	16.67%	41.67%	66.67%
Intensity	100%	0%	25%	75%

5.2. Evaluation of Noise types

We also grouped the data by noise type, as shown in table 2. As babble simulates the more natural Cocktail Party Effect [22], we might expect to see differences between how people communicate at difference noise levels. Each table grouped tokens together as described in the previous section, and then trends were visually identified, with patterns highlighted in the table. Babble noise in general shows positive changes. Height, area, and intensity have probabilities of 75%, 75%, and 66.67% of being greater in babble. This matches acoustic research by Panikos et al. [10].

However, white noise generally shows negative changes in noisy environments. Lip width has a probability of 66.67 percent to be narrower in noise, and interestingly, duration also has a probability of 83.33 percent of being shorter in noise. This is similar to the female group results. This result may be correlated to the background white noise, as research showed worsened performance for attentive children occurred while exposed to white noise [23]. Although this is only a pilot study, and should not be regarded as definitive, it does show that the type of noise does make a difference to how people react visually, as supported by research into acoustic only effects.

Table 2: *Parameter changes for babble and white noise environments, at all noise levels and both genders, showing grouped tokens and change in KDE, with majority changes highlighted.*

Parameter	Probability Change			
	Babble		White Noise	
	Positive	Negative	Positive	Negative
Duration	50%	50%	33.33%	66.67%
Width	33.33%	66.67%	16.67%	83.33%
Height	75%	25%	41.67%	58.33%
Area	75%	25%	50%	50%
Intensity	66.67%	33.33%	58.33%	41.67%

5.3. Evaluation of Noise levels

Table 3 shows the mean values of the different visual parameters, and also how the mean pixel values change for female (top) and male (bottom) speakers at different noise levels and noise types. The results show that mouth height and area tend to be larger in overall noise types compared to in no noise. The effect in babble noise tends have a slightly higher influence than white noise in both male and female groups, as discussed previously. However, the difference between different levels of noise (60-66dB vs 72-78 dB) are not significant. This suggests that noise has an effect, and the noise type also has an effect, but the precise noise level is not significant.

Table 3: *Mean values of female (top) and male (bottom) speaker parameters at different noise levels and different noise types, showing number of pixels, for duration, width, height, and area, and sum of pixel values for intensity.*

Mean values (Female)					
Parameter	Visual speech data				
	No Noise	Babble Noise 60-66 dB	Babble Noise 72-78 dB	White Noise 60-66 dB	White Noise 72-78 dB
Duration	11	11	10	10	10
Width	22	17	15	15	15
Height	12	14	14	12	12
Area	773	996	1057	909	909
Intensity	118048	130487	133255	118337	115503
Mean values (Male)					
Parameter	Visual speech data				
	No Noise	Babble Noise 60-66 dB	Babble Noise 72-78 dB	White Noise 60-66 dB	White Noise 72-78 dB
Duration	10	10	9	9	10
Width	23	24	24	23	23
Height	10	13	13	13	12
Area	723	904	999	920	922
Intensity	110934	164330	173148	152657	150149

6. Conclusions

Using a Gabor based visual feature extraction system, in this work, we recorded a pilot Lombard Speech database and used 3 male and 3 female speakers to investigate the Visual Lombard Effect, by extracting features of the mouth region, including lip area, width, height, and intensity. Generally, in noisy environments, male speakers tend to have a larger mouth area, agreeing with work in the literature. However, female speakers tended to have a smaller mouth area in noise, and their speech duration in noise tends to be shorter. Another key difference is how speakers react to noise types. There is an obvious and clear difference between silence and noise, which is to be expected, but also a clear difference between babble and white noise. This shows that the type of noise matters, and that people handle different noises differently, even at the same noise level. However, we also found that while there is a clear difference between noise and silence, the difference between noise levels is not significant. This is to be expected, considering the limited range of mouth opening possible, and that exaggerated speech is not helpful for natural communication. Overall, there are gender differences and noise type differences which should be taken into account when designing realistic speech recognition systems. The approach of testing lipreading systems by simply mixing acoustic noise and leaving the visual signal unchanged has very real world limitations.

7. Acknowledgements

This data was recorded as part of a Nuffield Trust Research Placement at the University of Stirling, Scotland. The authors would like to especially thank Dawn Hearsum for her hard work with data collection, and we would also like to thank Professor Roger Watt and Professor Leslie Smith for their advice.

8. References

[1] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[2] C. N. Hanley and D. G. Harvey, "Quantifying the Lombard effect," *Journal of Speech and Hearing Disorders*, vol. 30, no. 3, pp. 274–277, 1965.

[3] M. Fitzpatrick, J. Kim, and C. Davis, "The effect of seeing the interlocutor on auditory and visual speech production in noise," *Speech Communication*, vol. 74, pp. 37–51, 2015.

[4] J. Kim, C. Davis, G. Vignali, and H. Hill, "A visual concomitant of the lombard reflex." in *AVSP*, 2005, pp. 17–22.

[5] M. Garnier, M. Dohen, H. Lœvenbruck, P. Welby, and L. Bailly, "The lombard effect: a physiological reflex or a controlled intelligibility enhancement?" 2006.

[6] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics," *Speech communication*, vol. 20, no. 3, p. 255, 1996.

[7] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[8] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[9] M. Elhilali and S. Shamma, "Information-bearing components of speech intelligibility under babble-noise and bandlimiting distortions," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4205–4208.

[10] P. Heracleous, C. T. Ishi, M. Sato, H. Ishiguro, and N. Hagita, "Analysis of the visual lombard effect and automatic recognition experiments," *Computer Speech & Language*, vol. 27, no. 1, pp. 288–300, 2013.

[11] A. Abel and A. Hussain, "Novel two-stage audiovisual speech filtering in noisy environments," *Cognitive Computation*, pp. 1–18, 2013.

[12] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, vol. 27, no. 3, pp. 763–779, 2013.

[13] Y. Xu, Y. Li, and A. Abel, "Gabor based lipreading with a new audiovisual mandarin corpus," in *Advances in Brain Inspired Cognitive Systems*, J. Ren, A. Hussain, H. Zhao, K. Huang, J. Zheng, J. Cai, R. Chen, and Y. Xiao, Eds. Cham: Springer International Publishing, 2020, pp. 169–179.

[14] A. Abel, C. Gao, L. Smith, R. Watt, and A. Hussain, "Fast lip feature extraction using psychologically motivated gabor features," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2018, pp. 1033–1040.

[15] S. C. Dakin and R. J. Watt, "Biological "bar codes" in human faces." *Journal of vision*, vol. 9 4, pp. 2.1–10, 2009.

[16] D. King, "Dlib c++ library," *Access on: http://dlib. net*, 2012.

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[18] J.-C. Yen, F.-J. Chang, and S. Chang, "A new criterion for automatic multilevel thresholding," *IEEE Transactions on Image Processing*, vol. 4, no. 3, pp. 370–378, 1995.

[19] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lipreading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018.

[20] M. Piccardi and O. Perez, "Hidden markov models with kernel density estimation of emission probabilities and their use in activity recognition," pp. 1–8, 2007.

[21] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Lœvenbruck, "An acoustic and articulatory study of lombard speech: Global effects on the utterance," in *Ninth International Conference on Spoken Language Processing*, 2006.

[22] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[23] G. B. Söderlund, S. Sikström, J. M. Loftesnes, and E. J. Sonuga-Barke, "The effects of background white noise on memory performance in inattentive school children," *Behavioral and brain functions*, vol. 6, no. 1, p. 55, 2010.